Gazdasági hírek tartalmának feldolgozása banki előrejelző rendszer támogatásához

Tarczali Tünde, Skrop Adrienn, Mokcsay Ádám

Pannon Egyetem, Rendszer- és Számítástudományi Tanszék
8200 Veszprém, Egyetem u. 10.
{skrop,tarczali}@dcs.uni-pannon.hu
adam@mokcsay.hu

Kivonat: Kutatásunk célja egy olyan „early warning" mechanizmus és alkalmazás kifejlesztése, amely a weben megjelenő „szoft" információk feldolgozásán alapulva pénzügyi intézetek számára kockázat előrejelző szolgáltatást nyújt. A rendszer feladata a vizsgálandó alanyokkal kapcsolatos hírek, úgynevezett szoft információk keresése a weben, a talált hírek vizsgálata szövegbányászati eszközökkel, jellemzőik azonosítása és ezek alapján előre meghatározott kockázati kategóriákba sorolása. Cikkünkben ismertetjük a tervezett rendszer felépítését és az elkészült modulok működését.

1 Bevezetés

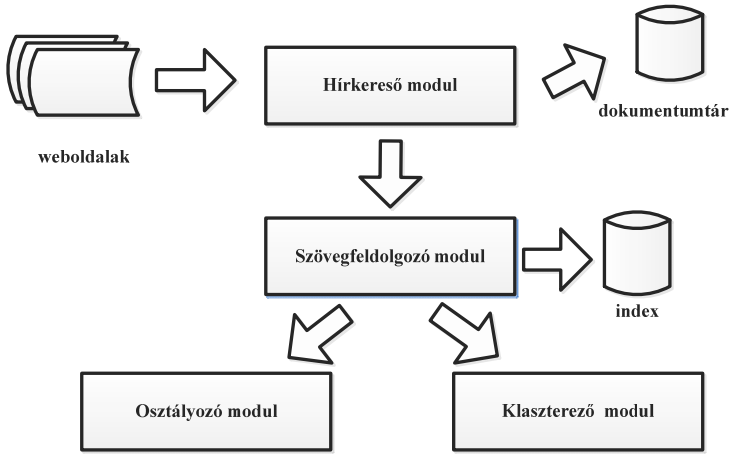
A hírelemzés szöveges hírek különböző kvalitatív és kvantitatív tulajdonságainak mérésével, elemzésével foglalkozik. Ilyen tulajdonságok például a szentiment, a relevancia és az újdonság. A hírelemzés magában foglalja mindazon technikákat és módszereket, melyek segítségével nyilvános információforrások feldolgozhatók, osztályozhatók [3]. A hírelemzés egyik fontos területe a gazdasági hírek elemzése, amely elsősorban azzal foglalkozik, hogy különböző gazdasági hírekre mikor és miként kell reagálnia a piacnak ahhoz, hogy a profitot növelni tudják.

A 2008 óta tartó és a pénzügyi szektort az ügyfeleik helyzetén keresztül is érintő pénzügyi-gazdasági válság középpontba helyezte a hitelkockázat minél hatékonyabb kezelésére és esetlegesen a kockázati tényezők előremutatására irányuló alkalmazások kifejlesztését. Jelen kutatás célja egy olyan automatizált kockázat előrejelző (early warning) módszer kifejlesztése, amely múltbéli információkból építkezve próbálja időben felismerni és jelezni az ügyfelek nem teljesítési kockázatát. A rendszer sajátossága, hogy a bankokban szokásos belső minősítésen alapuló módszer helyett az ügyfelek fizetésképtelenségre vonatkoztatott kockázatát a weben róluk megjelenő szoft információk szemantikai elemzésével jelzi előre.

2 A rendszer felépítése és működése

A tervezett rendszer felépítését az 1. ábra szemlélteti. A Hírkereső modul feladata a figyelendő alanyokra – ügyfelekre – vonatkozó, időzített keresések futtatása a weben.

A Hírkereső modul két funkciót lát el: egyrészt a múltbéli céges információk alapján mintákat gyűjt az osztályozáshoz használandó tanító minták meghatározásához, másrészt jelzi, ha egy ügyféllel kapcsolatban új hír jelent meg a weben. A weboldalak feldolgozását a Szövegfeldolgozó modul végzi. A modul feladata az internetes hírek előfeldolgozása, vektortér modellbeli reprezentálása [6], korpusz előállítása és a híreknek a tartalmazott szavak alapján történő kategorizálásának támogatása. A szemantikailag hasonló dokumentumok klaszterezését a Klaszterező modul végzi. A klaszterezés az AI^{2}R adaptív klaszterező eljárás segítségével történik [1]. A Hírkereső modul által szolgáltatott új hírek kockázati kategóriákba sorolása az Osztályozó modul feladata. Az osztályozásra naiv vektortér alapú módszert alkalmazunk [2], így a hasonlóság mértékének változtatása kevésbé számításigényes.



1. ábra. A rendszer felépítése.

2.1 Hírkereső modul

A Hírkereső modul feladata releváns, nem strukturált szoft információk keresése a weben. A modul megvalósítása hagyományos kulcsszavas metakeresővel történt. A metakereső olyan, webszervereken keresztül elérhető szoftver, mely egy adott kérdést elküld több webkeresőnek, összegyűjti és – valamilyen eljárással – egyesíti az eredményeket. A metakereső legfőbb előnye, hogy több kereső érhető el egyetlen, egyszerű interfésszel.

A megvalósított metakereső a Google és a Bing találati listáját használja fel, kezdőképernyőjét a 2. ábra mutatja. A metakereső egy weblapon keresztül érhető el, amelyet PHP motor generál, ezzel biztosítva annak dinamikus mivoltát, hiszen a működés során adatbázissal dolgozik a rendszer. Az adatbázist MySQL program kezeli, a rendszer pedig egy Linux alapú szerveren helyezkedik el. A felhasználó több paramétert képes megadni egy kereső kifejezés felvételénél, amelyet a program a beépít az egyes keresők felé intézett kérésbe. Ezekkel a paraméterekkel a keresés időzítése állítható be. Lehetőség van kereső-kifejezések importálására is, ebben az esetben egy XML kiterjesztésű fájlt vár a rendszer bemenetként.

2. ábra. Hírkereső modul kezdőoldal

A metakereső két alapvető tulajdonsága, hogy a keresés előre meghatározott kulcsszavak alapján történik, valamint a metakereső által visszaadott találati lista elemzése, a releváns oldalak végső ellenőrzése szakértő által történik. A Hírkereső találati listáját a 3. ábra szemlélteti.

3. ábra. Hírkereső modul találati lista

Minden előre definiált kereső kérdéshez meghatározásra kerül egy találati lista. A rendszer feladata, hogy a találati listát a beállított időzítésnek megfelelően frissítse és jelezze új, potenciálisan releváns találatok megjelenítését. A program lehetőséget biztosít az eredmények exportálásra, amely egy XML kiterjesztésű fájlt eredményez. A Kezelő menüpont segítségével a korábbi beállításokat módosíthatjuk.

2.2 Szövegfeldolgozó modul

A modul feladata az interneten fellelhető információk feldolgozása és gazdasági felszámolásra utaló releváns szavak kiemelése. Bemenetként a modulban megadhatóak hírekre mutató internetes linkek, vagy a Hírkereső modul által kimenetként szolgáltatott XML kiterjesztésű fájl, amely linkgyűjteményeket tartalmaz, akár meghatározott csoportokat is alkotva. Ennek segítségével egyszerre több, a szakértő által kiválasztott cikk együttes vizsgálatára nyílik lehetőség. A beolvasás lehetőségeit szemlélteti a 4. ábra.

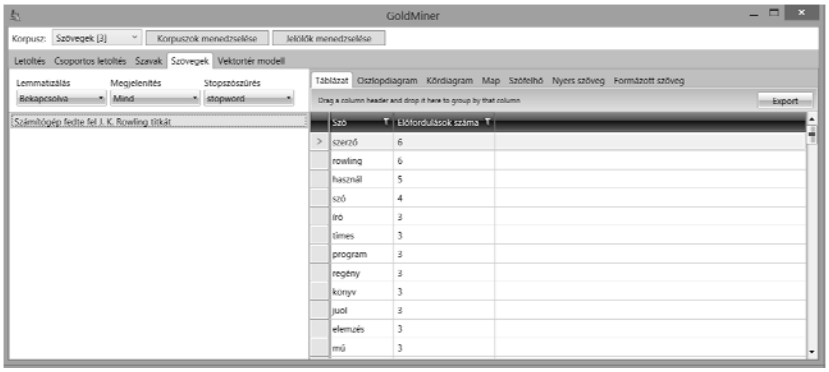
4. ábra. A hírek letöltése link megadásával

Az ábrán látható módon a cikkre mutató link megadásával a szoftver az internetről letölti a cikket és ezután történik meg annak feldolgozása. A hírek letöltésére automatizált letöltőket építettünk be a szövegelemző szoftverbe. Nem volt célunk saját letöltő készítése, hiszen a projekt céljának eléréséhez megfelelőek voltak a beépített automatikus letöltők. Egy linken található cikk betöltése mellett – a munka megkönnyítésére – lehetőség van több cikk egyidejű letöltésére is. Ennek megvalósítására egy XML file-t hoztunk létre, amely a következő formátumban tartalmazza a cikkek elérhetőségét:

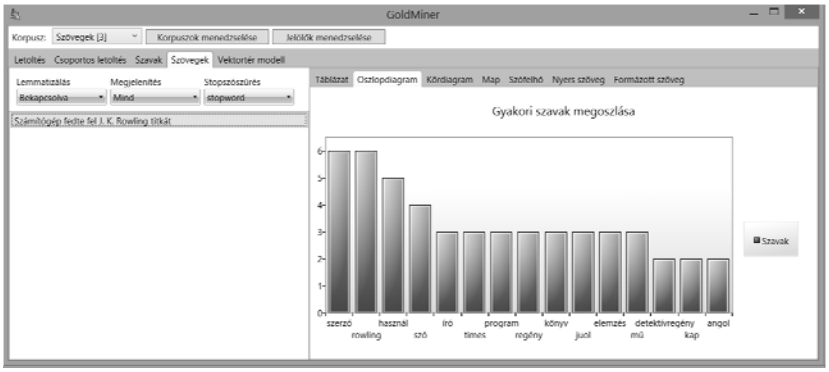
```
<?xml version="1.0" encoding="ISO-8859-2"?>
<download>

<corpus name="Első">
 <article href="http://link.url" />
 <article href="http://link.url2" />
 <article href="http://link.url3" />
</corpus>

<corpus name="Második">
 <article href="http://link.url4" />
 <article href="http://link.url5" />
 <article href="http://link.url6" />
</corpus>

<download>
```

A szövegek mondatokra, szavakra történő tagolásával (tokenizálással), valamint a stopszavak szűrésével végrehajthatóak olyan vizsgálatok, amely alapján a cikkekre vagy cikkgyűjteményekre jellemző szavakat, szóösszetételeket kaphatunk meg. A program beépített stopszótárral rendelkezik. A program első indításakor az adatbázis feltöltődik a stopszavak listájával. Ezekre a szavakra a program „Stopszó" címkét aggat. A stopszavak megadására külön listában van lehetőség, így a felhasználó maga is meghatározhatja ezeket. A projektünk témája indokolja, hogy jelen esetben arra keressünk választ, hogy az egyes cikkekben milyen szavak utalhatnak a vállalatok csőd közeli voltára. A programban lehetőség van a felhasználó által karban tartott jelölők tárolására, amelyekhez tetszőleges számú és nevű címke hozható létre. Ezen címkék hozzáadása történhet egy olyan szövegfájl alapján is, mely tartalmazza a jelölni kívánt szavakat.

Mivel a program a gazdaság képviselőinek készült, ezért szükség volt a statisztikai adatok grafikus megjelenítésére, amely segíti a szakértői értékelést. Erre mutatnak példát az 5-8. ábrák.



5. ábra. Táblázatos vizualizáció bekapcsolt lemmatizálás mellett

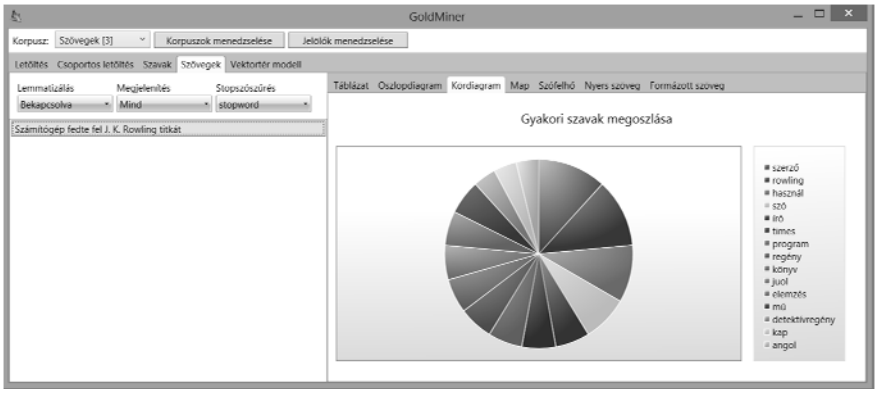


6. ábra. Cikkek vizuális elemzése oszlopdiagramon

Az egyenként történő statisztikai feldolgozás a tokenizálás után történhet a lemmák vizsgálatával, illetve anélkül. Itt egy beépülő modul segítségével vizsgáljuk a szavak

szótöveinek előfordulási gyakoriságát. A további feldolgozást segítendő lehetőség van a táblázat exportálására .xlsx formátumban.

A cikkekben található szavak előfordulási gyakoriságának elemzésére illetve a kimutatások elkészítéséhez a felület biztosít egy, az Excel programból jól ismert oszlopdiagramot, amely a vízszintes tengelyen tartalmazza a gyakran előforduló szavakat. Az oszlopok magassága a szó előfordulási gyakoriságával arányos.



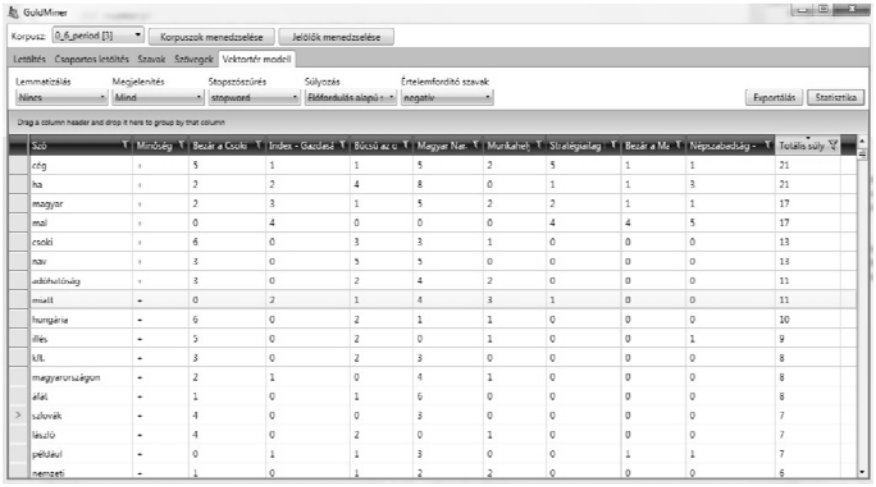
7. ábra. Cikkek vizuális elemzése kördiagramon

Az oszlopdiagram mellett az elemzés könnyítését szolgálja a kördiagrammal történő megjelenítés is. Itt a körcikkek vastagsága arányos a szavak előfordulási gyakoriságával.

8. ábra. Cikkek vizuális elemzése szófelhővel

Egy érdekes vizuális megjelenítést célzó ábra a szófelhő. Az interneten a cikkek megjelölésére gyakran használt eszköz a címkézés. A címkék előfordulásának gyakoriságát illetve a cikkek olvasásának gyakoriságát gyakran mutatják címkefelhővel. Ezt a megjelenítési módszert alkalmaztuk a szavak gyakoriságának bemutatására. A nagyobb betűvel megjelenő szavak jelentik a szövegben gyakran előforduló szavakat.

A program a szövegeket szöveggyűjteményekben, korpuszokban tárolja. A szoftver a karbantartott korpuszokból képes vektortér modell előállítására. Ennek szükségességét az adja, hogy a cikkek elemzése cégekhez és a tanító fázisban a csőd közeli állapothoz viszonyított időszakokra vonatkoztatva történik. Az alábbi kép mutatja a program működésének azt a fázisát, ahol egy cikkcsoportra vizsgáljuk a szavak előfordulását.



9. ábra. Vektortér modell kialakítása a kiválasztott cikkekre

A vektortér modellben [5] mindazon lehetőségek megvannak, amelyek az egyes cikkek elemzésénél is segítségünkre lehetnek. A kanonikus alak megtalálására alkalmazható eljárás például a szavak csonkolása. Ekkor szótőként általában nem a szótári szóalakot kapjuk, ám a legtöbb esetben ez is kellően pontos. Léteznek egyéb szótár alapú algoritmusok is. Ilyen algoritmus pl. a Porter féle algoritmus, Lovins-tövező, vagy a Snowball alapú magyar tövező. A szótövezést a Hunstem program végzi [4]. A program felismeri a szavak töveit, ezzel lehetővé téve a szótő szerinti csoportosítást és a generált vektortér modell dimenziószámának redukálását. A szavak szótövezése mellett megvalósításra kerültek olyan súlyozások, amelyek a különböző vizsgálatokat segítik. A következő súlyozási módszereket [6] implementáltuk:

- bináris
- előfordulás alapú
- logaritmikus
- gyakoriság alapú
- TF-IDF

Ezeken kívül lehetőség van az értelemfordító szavak vizsgálatára is. Két szó távolságban vesszük figyelembe azt, hogy a cikkekben megjelenő értelemfordító szavak negatív értelmet adnak egyes kifejezéseknek.

2.3 Klaszterező modul

A szemantikailag hasonló dokumentumok klaszterezését a Klaszterező modul végzi. Klaszterezés során a dokumentumokat – általában – diszjunkt halmazokba csoportosítjuk. Minden klaszter – bizonyos értelemben – hasonló dokumentumokból áll. A modul célja az azonos kockázati kategóriát képviselő hírek egy csoportba sorolása.

A különböző klaszterezési technikák közül a gazdasági területet igényeihez leginkább illeszkedő módszert kellett meghatározni. Az a fontos igény került figyelembevételre, hogy ne csak az azonos kifejezéseket tartalmazó cikkek, hanem egy cikkhez szemantikailag hasonló tartalmúak is egy klaszterbe kerüljenek. Ez az elvárás indokolta, hogy az interakciós információ-visszakereső I^2R (Interaction Information Retrieval) technikát választottuk.

Az I^2R matematikai modellje a mesterséges neuronhálózat alapvető állapotegyenletén alapszik. Eszerint a dokumentumok azonosíthatóak egy neuronhálózattal, ahol az egyes dokumentumok egy-egy neuronnak felelnek meg, amelyek képesek különböző szintű aktivitást produkálni. Egy új dokumentum szintén egy neuronnak felel meg, amely beépül a hálózatba – mint egy új objektum – és így a hálózat részlegesen megváltozik: új kapcsolatok alakulnak ki az új és az eredeti objektumok között, továbbá az eredeti hálózatban kialakult kapcsolatok egy része módosulhat. Ez a hatás indítja el a klaszterezési folyamatot.

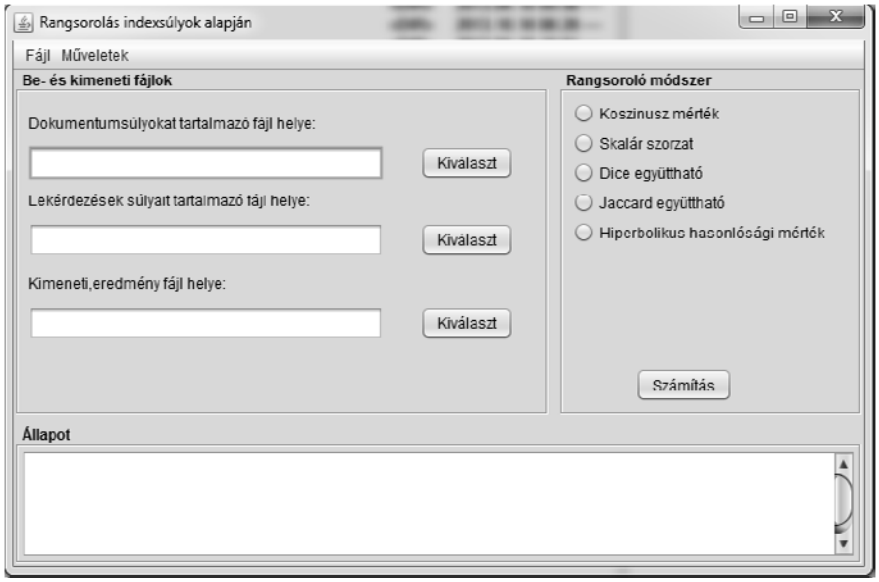
2.4 Osztályozó modul

A Hírkereső modul által szolgáltatott új hírek kockázati kategóriákba sorolása az Osztályozó modul feladata. A Klaszterező modul által meghatározott csoportok nem jellemezhetők a hagyományos értelemben vett címkékkel, hanem kockázati kategóriákat jelölnek, ezért az osztályozásra naiv vektortér alapú módszert alkalmazunk.

Mind a klaszterekben szereplő cikkeket, mind az új híreket a szentiment elemzés során definiált vektortérbeli vektorokként ábrázoljuk. Az új hírek klaszterbe sorolása a vektortérben használt hasonlósági mérték segítségével történik. A módszer azon alapul, hogy az új hírt reprezentáló vektor és egy klaszterbeli vektor elég közel vannak-e egymáshoz. A vektorok hasonlóságát különböző hasonlósági mértékek segítségével lehet mérni.

A vektortér modellt hagyományosan euklideszi térben definiálják. Az Osztályozó modulban implementálásra kerültek az euklideszi tér szokásos hasonlósági mértékei, mint a belső szorzat, a koszinusz mérték, a Dice együttható és a Jaccard együttható. A hagyományos modell mellett implementálásra került a hiperbolikus információ-visszakereső modell is, melynek lényege, hogy a benne alkalmazott hasonlósági

mérték a Cayley-Klein hiperbolikus távolságból származik. Gyakorlati tesztsorozatok segítségével fogjuk meghatározni, hogy melyik módszer alkalmas gazdasági hírek osztályozására. Az osztályozó modult a 10. ábra szemlélteti.



10. ábra. Osztályozó modul

3 A kutatás eredményei

Kutatás-fejlesztési feladatunk célja az interneten elérhető gazdasági tartalmú információk, hírek megkeresése és feldolgozása, a releváns tartalom kinyerése és a cikkek osztályozása. A kutatás első lépéseként meghatározásra kerültek azok a jellemzően szöveges információk, amelyek valamely negatív esemény bekövetkezését jelezhetik. A múltbéli céges információk elemzésére a kutatáshoz rendelkezésre áll a Dun&Bradstreet teljes magyar sokaságra vonatkozó minta adatbázisa. Szakértői segítséggel kiválasztásra kerültek azok cégek, illetve ezután azok a rájuk vonatkozó cikkek és időszakok, amelyek elemzése a készített alkalmazással folyamatosan történik. A meghatározott információk alapján lefolytattuk azokat az internetes kereséseket, amelyek alapján a cikkek szakértők általi szűrésével előállt az a releváns információkat tartalmazó cikkhalmaz, amelynek feldolgozásával a csőd előrejelzése támogatható. Ezen adatok alapján webes kereséssel felállítjuk azon tanító halmazokat, amelyek alkalmazásával a megjelenő cikkekről eldönthető, szolgáltatnak-e információkat a cégek pénzügyi helyzetével kapcsolatban.

Köszönetnyilvánítás

A publikáció az Európai Unió, Magyarország és az Európai Szociális Alap társfinanszírozása által biztosított forrásból a TÁMOP-4.2.2.C-11/1/KONV-2012-0004 azonosítójú „Nemzeti kutatóközpont fejlett infokommunikációs technológiák kidolgozására és piaci bevezetésére” című projekt támogatásával jött létre.

A kutatás a GOP-1.1.1-11-2011-0045 azonosítójú EWS – Adat- és folyamatbányászati algoritmusokon alapuló automatizált kockázat előrejelző rendszer prototípusának fejlesztése pénzügyi intézetek számára című projekt támogatásával valósult meg. A cikk tartalma kizárólag a szerzők felelőssége, és nem feltétlenül tükrözi a támogatók álláspontját.

Hivatkozások

1. Dominich, S.: Connectionist interaction information retrieval. Information processing & management. Vol. 39(2) (2003) 167–193.
2. Góth, J., Skrop, A.:Varying retrieval categoricity using hyperbolic geometry. Information Retrieval. Vol. 8(2) (2005) 265–283
3. Mitra, G., Mitra, L.: The Handbook of News Analytics in Finance. John Wiley & Sons (2011)
4. Németh, L.: A Szószablya fejlesztés. 5th Hungarian Linux Conference (2003)
5. Subecz, Z.: Információkinyerés természetes nyelvű szövegekből. Szolnoki Tudományos Közlemények XV., Szolnok (2011)
6. Tikk, D. (szerk.): Szövegbányászat. Az informatika alkalmazásai sorozat. ISBN 978-963-9664-45-6. (2007)